DOCUMENT RESUME

ED 174 661 TM 009 537

AUTHOR Massey, Randy H.; And Others
TITLE Performance Appraisal Ratings: The Content Issue. Final Report, June 1976 through August 1978.
INSTITUTION Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-78-69
PUB DATE Dec 78
NOTE 21p.
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock Number 671-056/109)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Content Analysis; Evaluation Criteria; Individual Characteristics; *Officer Personnel; *Peer Evaluation; *Personnel Evaluation; *Rating Scales; *Task Performance; Test Items; Test Reliability; *Work Attitudes
IDENTIFIERS Air Force

ABSTRACT

Three kinds of rating statements, trait-oriented, worker-oriented, and task-oriented were evaluated in a context permitting the comparisons to be made in terms of criteria external to the ratings. One hundred twenty Air Force noncommissioned officers assigned to seminar groups of 13 or 14 were involved. No evidence of superiority was found for any of the three sets although significant correlations with various external criteria were obtained in all three experimental conditions. Significant differences were also found among the three rating sub-groups comprising each of the three treatment groups although these rating sub-groups were assigned randomly to the three treatment groups. The importance of controlling for group effects in peer group studies was noted. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-78-69

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AIR FORCE HUMAN RESOURCES LABORATORY

ED174661

IM009 537

PERFORMANCE APPRAISAL RATINGS:
THE CONTENT ISSUE

By

Randy H. Massey, Capt, USAF
Cecil J. Mullins
James A. Earles

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

December 1978
Final Report for Period June 1976 – August 1978

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, under project 2313, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Randy H. Massey (PEP) was the Principal Investigator for the Laboratory.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

LELAND D. BROKAW, Technical Director
Personnel Research Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-78-69	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: PERFORMANCE APPRAISAL RATINGS: THE CONTENT ISSUE	5. TYPE OF REPORT & PERIOD COVERED: Final, June 1976 – August 1978
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Randy H. Massey, Cecil J. Mullins, James A. Earles	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Personnel Research Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F, 2313T602
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235	12. REPORT DATE: December 1978
	13. NUMBER OF PAGES: 20
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

SM Study Nr. 6728

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

rating statements
performance ratings
peer ratings
peer rankings
task-oriented dimensions
trait-oriented dimensions
worker-oriented dimensions

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Three kinds of rating statements, trait-oriented, worker-oriented, and task-oriented, were evaluated in a context permitting the comparisons to be made in terms of criteria external to the ratings. One hundred twenty Air Force noncommissioned officers assigned to seminar groups of 13 or 14 at the Air Training Command NCO Academy, Lackland AFB Annex, were involved in this experiment.

No evidence of superiority was found for any of the three sets although significant correlations with various external criteria were obtained in all three experimental conditions. Significant differences were also found among the three rating sub-groups comprising each of the three treatment groups although these rating sub-groups were assigned randomly to the three treatment groups. The importance of controlling for group effects in peer group studies was noted.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

PREFACE

This research was conducted under project 2313, Research on Human Factors in Aero Systems; task 2313T6, Force Acquisition, Assignment, and Evaluation.

TABLE OF CONTENTS

		Page
I.	Introduction	5
II.	Method	7
	Sample	7
	Rating Scales	7
	Rating Tasks	7
	Research Approach and Rationale	7
	Data Analysis	8
III.	Results and Discussion	10
IV.	Summary and Conclusions	13
	References	13
	Appendix A: Rating Dimensions	15

LIST OF TABLES

Table		Page
1	Analysis of Variance of Number of "Hits" (Correct Profile Identifications) by Treatment and Seminar Group	10
2	Number of Profile Identifications (Hits) by Treatment and Seminar Group	10
3	Rank Order Correlations Among Unidentified Profile Rankings, Peer Rankings, and Official Rank by Treatment and by Seminar Group	11
4	Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Peer Rankings by Treatment and by Seminar Group	11
5	Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Official Rankings by Treatment and by Seminar Group	11
6	Analysis of Variance of Squared Deviations between Peer Rankings and Official Rankings by Treatment and by Seminar Group	12

PERFORMANCE APPRAISAL RATINGS: THE CONTENT ISSUE

I. INTRODUCTION

Much research done on ratings has been concerned with efforts to determine the best stimulus statements to use in a rating situation. Unfortunately, in much of this research "best" has been defined in terms of psychometric properties inherent in the ratings. Little research has been done employing external criteria for evaluating rating statements.

This is one in a series of studies intended to help resolve the content issue of rating statements. This study focuses on the relative merits of rating statements with content selected to represent different points on a continuum from highly job-specific statements to person-oriented, trait-like statements. A context was constructed which provides an opportunity to evaluate the usefulness of various sets of rating statements against criteria external to the ratings, rather than the more traditional method of evaluating rating statements in terms of their internal psychometric characteristics.

The generally accepted viewpoint is that the more specific observable behaviors are more accurately rated than general personality descriptive statements. This viewpoint appears to be based more on the selective appraisal of a narrow spectrum of studies than on an appraisal of all studies conducted in the field (Kavanagh, 1971). In any case, the difficulties and controversial issues inherent in ratings have been well documented (e.g., Barrett, 1966; Kavanagh, 1971; Ronan & Prien, 1971; Schmidt & Kaplan, 1971).

A popular scaling procedure designed to measure job performance is the Behavioral Expectation Scales (BES) developed by Smith and Kendall (1963). In this procedure, the important performance dimensions are identified and defined by a group of individuals responsible for evaluations. The scales are anchored by actual job behaviors which represent specific performance levels. The BES has had considerable intuitive appeal, and there have been many proponents of the technique (e.g., Campbell, Dunnette, Arvey, Hellervik, 1973; Campbell, Dunnette, Lawler, & Weick, 1970; Dunnette, 1966; Landy, Farr, Saal, & Freytag, 1976; Zedeck & Blood, 1974). BES scales have also been developed for a variety of occupations (e.g., Arvey & Hoyle, 1974; Landy et al., 1976; Smith & Kendall, 1963). This may account for the belief that behavior-based rating statements are superior to trait-oriented statements.

Despite its popularity, a review of studies in which BES was compared to other formats does not provide overwhelming support for the BES. Burnaska and Hollmann (1974), in examining the psychometric characteristics of three different rating scale formats (BES, BES without anchors, and another set of a priori dimensions), found no differences among the formats with respect to halo, rater bias, or leniency. They concluded that "There is no evidence for superiority of any one format." Other investigators (e.g., Dickinson & Tice, 1973; Zedeck & Baker, 1972) have found little advantage in terms of discriminant or convergent validity of BES obtained ratings. BES ratings have also been found to be non-transferable within the same occupation from the original developed setting to another similar work setting (Borman & Vallon, 1974). The BES exhibited no superiority over a more simple scale (BES without anchors) on interrater agreement and halo effect. In fact, the simpler scale showed significantly less leniency effect (lower adjusted mean ratings and greater adjusted standard deviations) than the BES format. In short, the literature does not provide overwhelming support for the superiority of BES over other scale formats.

Other popular methodologies include deriving rating scales based on patterns of job requirements (McCormick, 1959) and the multitrait-multimethod approach to measuring job performance (Campbell & Fiske, 1959). McCormick (1959) emphasizes the importance of using job-oriented and worker-oriented statements primarily derived from job analysis techniques. Job-oriented statements describe the job content, or what is accomplished by the worker (repair water pump, inspect lubrication system, drive pickup truck, etc.). Worker-oriented statements tend to characterize generalized human behaviors or worker

characteristics which are usually descriptive across many different jobs (observe visual displays, judge condition or quality, manually pour ingredients into container, etc.). In the multitrait-multimethod approach, data from many traits and raters are analyzed for convergent and discriminant validity (Campbell & Fiske, 1959). The concepts of convergent and discriminant validity, in the context of the Campbell and Fiske paper, appear to apply primarily in situations where there is no clearly preferred single target or criterion variable available. Convergent validity is represented by the size of the correlations among data sets from independent sources, such as separate raters, and discriminant validity is represented by the size of the correlations among different variables obtained from the same source, such as separate rating statements from the same rater. Obviously, one prefers the convergent validity correlation coefficients to be high and the discriminant validity correlation coefficients to be low. In the rating situation, to the degree that correlation coefficients representing discriminant validities are high, one suspects that a large amount of halo error is present. The multitrait-multimethod approach offers evidence that traits can be effective in performance evaluation devices (Kavanagh, 1971; Kelley & Fiske, 1951). The BES and McCormick (1959) approaches basically assume the superiority of behavior-based or task-oriented type dimensions. There is no comparative evidence to indicate the superiority of any of the methodologies.

A common issue underlying all rating methodological approaches is the "content issue" defined by Kavanagh (1971) as "the issue of the relative representativeness of traits . . . along a continuum ranging from subjective to objective, abstract to concrete, or personality to performance." He concluded that there is no overwhelming evidence to indicate the superiority of behavior-based over trait-oriented dimensions. He further suggests that contradictory findings across reliability and validity studies could be partially attributed to a failure to resolve or control for the "content issue." Resolution of this issue may give insight into the effectiveness of various performance evaluation methodologies, particularly in relation to time and cost expended. Settlement of this issue can also have significant explanatory value accounting for the numerous contradictory findings that exist in performance appraisal research.

Kavanagh, MacKinney, and Wollins (1971) were the first to directly address the content issue, using the multirater-multimethod approach, by investigating middle managers using performance ratings from superiors and two subordinates. They found more convergent validity for personal traits than performance traits, but no difference for discriminant validity. Although the higher personal trait convergent validity was accompanied by a greater degree of "halo," the overall conclusion was that ratings of personal traits did as well as the ratings of performance traits.

Since Kavanagh (1971), the content issue has been almost entirely ignored. Recently Borman and Dunnette (1975) attempted to resolve the content issue by comparing behavior-based statements with trait-oriented statements. Their conclusions were, "at present little empirical evidence exists supporting the incremental validity of performance ratings made using behavior scales." Unfortunately, there are methodological problems associated with their study. They compared three different rating systems (performance anchored, performance non-anchored, and trait-oriented statements obtained from the Naval Officer Fitness Report), rather than just comparing three rating formats. In sum, the study did not directly focus on the content issue of rating criteria, but rather on the effectiveness of three different rating systems. Among other experimental difficulties, they compared different numbers of rating statements between treatments and included trait-like statements (integrity, responsibility, and dedication) within the performance treatment category.

It seems clear, then, that the issue of the preferred content for rating statements has in no way been resolved by previous research. This study is one in a series of studies using criteria external to the ratings to attempt such a resolution. It is anticipated that this approach will be more effective in resolving the content issue than were past studies that employed internal characteristics of the rating instrument as criteria for judging the excellence of rating statements.

II. METHOD

Sample

One hundred twenty students assigned to the Air Training Command (ATC) NCO Academy at Lackland AFB Annex completed the rating tasks. The study included nine separate seminar groups, each consisting of 13 or 14 noncommissioned officers (E6s to E7s) whose length of military service was 10 to 17 years.

Rating Scales

The treatment conditions in this study varied across three different types of rating statements (task-oriented, worker-oriented, and trait-oriented). Ten rating statements representing each of the three different kinds of rating content were included in the study. These were determined by consultation with instructors, administrative officials, and students. Previously conducted studies were also reviewed to identify factors. Each of the 10 rating attributes was rated on a 5-point scale as follows:

Specific Ratable Attribute	Below Average	Average	Above Average	Well Above Average	Outstanding

Trait-oriented attributes also included a brief descriptive definition. See Appendix A for a complete list and description of the rating statements.

Rating Tasks

The research was conducted in two phases. In Phase I, each student rated all members in his seminar group on one and only one of the three different types of statements—task-oriented, worker-oriented, and trait-oriented. This phase resulted in the generation of individual profiles based on the group's evaluation of each member on each of the 10 selected rating attributes.

In Phase II, about 2 weeks later, the experimenter handed out the profiles to the seminar group without an identifying name on the profiles. Each subject was required to perform three tasks: first, he had to rank-order the profiles according to predicted seminar class rank; second, he had to identify to whom each profile belonged; and third, he had to predict the final school seminar class rank of his seminar peers without any regard to profile considerations. Subjects appeared unaware of the nature of the study until Phase II research when they were asked to identify each of the profiles.

Research Approach and Rationale

Many studies into the relative efficiency of sets of rating statements have apparently started with a basic set of assumptions. First, the raters are subject to leniency error resulting in elevated means and to halo error revealed by small standard deviations among the ratings assigned. Since these two forms of rating error are revealed by the indicated statistics, a study of means and standard deviations forms a basis for comparison among sets of rating statements which may be used to distinguish among sets as to their goodness. Second, if rating statements are meaningful, and if raters are accurate in their perceptions of ratees, then inter-judge agreement, in the form of correlations among sets of ratings issuing from different judges, will be an expression of the goodness of a set of ratings. Third, the most useful way to compare sets of rating statements with each other lies in the comparisons which can be made among the summary statistics produced by the ratings. If one accepts these assumptions, then it follows that the best way to compare sets of rating descriptions is as it has frequently been done—the best set is that set which produces lower means, larger standard deviations, and larger inter-judge correlation coefficients.

However, the foregoing assumptions are subject to challenge. Taking them in order:

1. The evidence seems clear that leniency and halo errors do occur. It is less clear how important these two errors are in a family of other possible errors (e.g., racial bias, low rater motivation, low observability of the ratee, and others). It is also clear that there is not a direct relationship between leniency error and larger means or between halo error and smaller standard deviations. A person who is good on one dimension is more likely also to be good on whatever other dimensions are being considered. This is true whether the "goodness" metric is derived from ratings, from tests, or from any other reasonable source. Therefore, some portion of "halo error" may reflect true conditions, and be no error at all.

2. Inter-judge agreement *may* sometimes be a sufficient basis for comparing sets of rating statements, but it is not unusual for groups of judges to agree on a decision which additional facts show to be in error. If one may postulate individual differences among raters in respect to their ability to perceive ratees accurately, which seems plausible, then one must agree that some raters will provide better ratings. If some raters are better than others, it seems naive to expect that their ratings of a given characteristic will fall eternally at the mean of ratings given on that characteristic.

3. In this study, an approach is taken which provides a better basis for making comparisons across rating sets than does the traditional psychometric comparison. The approach is constructed around the concept of "hits"; that is, the number of times a rater can correctly identify anonymous profiles of his peers, constructed around various sets of descriptor statements.

If a rating statement is useful in describing a person, and if a group of raters can agree to some extent on the elevation of this characteristic in a ratee, then a profile of this ratee produced from a set of such statements should be identifiable as a rating "picture" of that individual. If a group of raters can recognize the individuals whom their profiles describe, then it seems more likely that the set of profiled characteristics can be useful in evaluating or predicting the performance of those individuals. The number of "hits" (correctly labeled profiles) should be useful in comparing one set of rating descriptions with another.

One analysis was made using hits as the dependent variable. The number of hits, however, at least in prior research (Curtis, Ratliff, & Mullins, 1977), has proved so small that something more sensitive was needed. A rater could conceivably misidentify the first profile considered; and that misidentification could cause him to miss the rest, even if only by a small margin—or he could be so insensitive to personal differences that he makes guess errors in all the identifications. The search for a sensitive measure of profile identification led to the use of the rank-order correlation as a possibly more effective measure of identification of peers than the simple count of correct identifications.

If a rater trying to identify anonymous profiles of his peers is confronted with 15 profiles, three of which have been rated very high on a particular characteristic, and if he believes correctly that peers B, H, and J are the three in his peer group highest on this characteristic, he may not know which of the three is peer B. He might specifically misidentify all three profiles, even though he has been correct in believing that these three profiles, as a set, represent peers B, H, and J. Although he has come close, his number of exact identifications, or hits, among these three profiles would be zero, no better than it would be for some less astute rater who believed B, H, and J were the lowest three in the peer group on that characteristic. In short, the "hits" measure contains no provision for crediting near misses, but the correlation between the ranking of unidentified profiles and the ranking of his named peers on the success dimensions should provide a continuum which the raw "hits" metric does not possess. A rank-order correlation between these two ranks should provide a sensitive measure of recognition far more powerful than the simple count of matched profiles.

Data Analysis

In order to apply the metric described in the preceding paragraph, three rankings were collected. First, an official ranking (OR) of the students, performed by the school, was available. Second, a ranking of the anonymous profiles (UP) was collected. Finally, a ranking of seminar members by their peers (PR) was

collected. This ranking was made using only a list of peer names, not profiles, and was made according to predictions of success in training.

The UP and PR rankings were group average ranks derived by summing all of the assigned ranks for each person in his seminar group, then converting that total sum of ranks back to a rank order ranging from 1 to 13 or 14 depending on the seminar's group size. These average ranks, UP and PR, represented a group conscensus on the perception of each seminar member by the group. The Official Class Rank (OR) was determined by class standing on four exams (312 points), drill evaluation (25 points), student evaluation (25 points), and communication skills (38 points).

Rank-order correlations for each rater were computed for the following purposes:

1. Correlation between unidentified profile ranking and named peer rankings (UP-PR)—One correlation coefficient was computed for each rater and was viewed as a more sensitive measure of hits than is the number of exact identifications of unlabeled profiles. This produced a new variable, the logic of which was explained above.

2. Correlation between unidentified profile rankings and official class rank (UP-OR)—One correlation coefficient for each rater. This variable indicates how well the rater can evaluate the operational criterion (OR) in terms of the statements available. Differences in effectiveness among the statement sets should be revealed in differences between the sizes of the average correlation coefficients. Average correlation coefficients across groups could have been computed by summing the numerators in the rho formula ($6\Sigma d^2$) and dividing by the sum of the denominators [$N(N^2 - 1)$]. The squared deviations (d^2) were used in the analysis of variance (ANOVA) since in this instance it provided a simpler and more accurate measurement variable in examining rank order effect than did the correlation coefficients themselves.

3. Correlation between names peer rankings and official class rank (PR-OR)—One for each rater. The average of this correlation coefficient would normally indicate the efficiency of peer ratings in predicting a criterion. In this case, however, there was considerable evidence that most of the subjects were well aware through intra-group discussion of how their peers had done on previous tests and were consequently aware of how they stood on the overall class evaluation. In short, they were ranking on direct information about their peers rather than on judgment based on indirect knowledge.

The primary analysis included testing to see if significant differences existed in terms of hits and the other dependent variables among the three treatment conditions. Since each seminar group was randomly assigned to one of the three treatment conditions, the experimental design resulted in the nesting of three seminar groups under each treatment conditions. The hierarchical design (Nested Factors) is usually used to test the effects among a number of treatments in certain types of experimental situations (Winer, 1962). Typical examples include investigating drug effects among a number of hospitals, studying teaching methods among a number of schools, or studying training methods among different individuals.

The hierarchical ANOVA is an efficient method of studying such experimental situations because it avoids multiple t-tests or non-orthogonal comparisons (Hays, 1963). The two-way hierarchical ANOVA in this experiment is also a more powerful statistical test than a one-way ANOVA that only tests for treatment effects, ignoring any group effects. In this design, the nested factors are controlled by statistical procedures. In many experimental situations, it is dangerous to assume that certain nested factors have no significant influence on treatment effects.

Two sources of variation were observed in the experimental data. The treatment effect was of primary interest, whereas the seminar group affiliation was of secondary interest. The null hypothesis, i.e., no difference between treatment means, was tested for both investigated sources of variation. The analysis of both sources of variation was accomplished by performing a two-way hierarchical ANOVA for experiments with unequal cell sizes, using the least-squares procedural method described by Timm and Carlson (1975).

"Hits" and the sum of the squared differences between UP and PR rankings, UP and OR rankings, and PR and OR rankings were the dependent variables used in the ANOVA analysis to determine whether

significant differences existed among treatment conditions. The squared differences between rank orderings were used rather than the rank-order correlations since the squared differences provided a simpler and more accurate measurement variable in examining rank order similarity.

III. RESULTS AND DISCUSSION

The hierarchical ANOVA summary for "hits," or correct identification of profiles is shown in Table 1. As expected, the "hit" measurement variable showed no significant differences among treatments. In essence, the rating "picture" for each individual produced by the three different sets of rating statements were equal in their descriptive power. However, seminar group effects within treatments were significant at the .01 level (Table 1). Table 2 shows the summary results of hits for seminar groups within treatments.

Table 1. **Analysis of Variance of Number of "Hits" (Correct Profile Identifications) by Treatment and Seminar Group**

Source	Sum of Squares	df	Mean Square	F
Treatment	6.215	2	3.107	.349
Seminar Groups Within Treatments	53.421	6	3.903	3.247*
Error (Within Groups)	304.379	111	2.742	

*Significant at .01 level.

Table 2. **Number of Profile Identifications (Hits) by Treatment and by Seminar Group**

Results	Treatment 1 (Seminar Group)			Treatment 2 (Seminar Group)			Treatment 3 (Seminar Group)		
	F	I	A	C	E	H	B	D	G
Group									
Total N	13	14	13	14	13	13	13	13	14
Total Hits	24	47	45	32	26	48	23	29	42
Mean Hits	1.86	3.36	3.46	2.29	2.00	3.69	1.77	2.23	3.00
SD Hits	1.63	1.82	2.37	1.90	1.68	1.55	1.30	1.30	.96
Treatment									
Total N		40			40			40	
Total Hits		116			106			94	
Mean Hits		2.90			2.65			2.35	
SD Hits		2.05			1.83			1.27	
T-Ratios									
Treatments 1 vs. 2 Comparison					$t = .574^{ns}$				
Treatments 1 vs. 3 Comparison					$t = 1.44^{ns}$				
Treatments 2 vs. 3 Comparison					$t = .85^{ns}$				

^{ns} = not significant.

The average rank-order correlations between the pairs of rankings appear in Table 3. Using Ferguson's (1966) table of significance for Spearman rhos, 25 of the possible 27 rhos were significant at the .05 level. Furthermore, most of the nine correlations possible in each treatment group were significant at the .01 level 21 in all), and only one correlation in each of treatments II and III was not significant. All correlations

Table 3. **Rank Order Correlations Among Unidentified Profile Rankings, Peer Rankings, and Official Rank by Treatment and by Seminar Group**

Rank Order Comparisons	Treatments								
	I (Worker) Seminar Groups			II (Task) Seminar Groups			III (Trait) Seminar Groups		
	F	I	A	C	E	H	B	D	G
UP and PR	.58*	.86**	.87**	.85**	.86**	.90**	.70**	.90**	.71**
UP and OR	.52*	.71**	.82**	.43	.65*	.85**	.37	.72**	.70**
PR and OR	.87**	.93**	.97**	.57*	.79**	.94**	.74**	.79**	.97**
Total N	13	14	13	14	13	13	13	13	14

Note. Critical values of rho, the Spearman rank correlation, were obtained from Ferguson (1966), Table G, p. 414.

*Significant at .05 level.
**Significant at .01 level.

demonstrated a similar pattern of significance in each of the three treatment conditions. The three rank order comparisons showed a high degree of agreement. This data analysis suggested that no one type of rating statement was superior for use in performance appraisal instruments. The purpose of these rank-order comparisons was to see whether the pattern of significance under each treatment was generally similar or different. However, the most definitive test for determining differences between treatments was the hierarchical ANOVA analysis.

Tables 4 to 6 show the hierarchical ANOVA summary for comparison of the rating statement treatment conditions with respect to the squared difference between the following rank-order comparisons: UP-PR, UP-OR, and PR-OR. The ANOVA results showed no significant difference between treatment conditions as reflected by the squared differences between the UP-PR rankings (viewed as a more sensitive measure of identification of unlabeled profiles), the UP-PR rankings (which indicate how well the rater can evaluate the operational criterion in terms of given stimulus statements), and the PR-OR rankings (normally indicating the efficiency of peer ratings in predicting a criterion).

Table 4. **Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Peer Rankings by Treatment and by Seminar Group**

Source	Sum of Squares	df	Mean Square	F
Treatment	9396.114	2	4698.057	.117
Seminar Groups Within Treatments	241470.876	6	40245.146	4.722*
Error (Within Groups)	945985.099	111	8522.388	

*Significant at .01 level.

Table 5. **Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Official Rankings by Treatment and by Seminar Group**

Source	Sum of Squares	df	Mean Square	F
Treatment	12127.327	2	6063.663	.0922
Seminar Groups Within Treatments	394330.700	6	65721.783	13.051*
Error (Within Groups)	558976.730	111	5035.836	

*Significant at .01 level.

Table 6. **Analysis of Variance of Squared Deviations between Peer Rankings and Official Rankings by Treatment and by Seminar Group**

Source	Sum of Squares	df	Mean Squares	F
Treatments	119253.060	2	59631.530	.769
Seminar Groups Within Treatments	465196.015	6	77532.668	16.160*
Error (Within Groups)	532553.566	111	4797.780	

*Significant at .01 level.

The PR-OR rank order coefficient, however, cannot be considered an unbiased indicator since there was considerable evidence that most subjects were ranking on information based on knowledge of test performance acquired through intra-group association, rather than judgment based solely on observation of peer activities and traits.

Although no significant rank-order differences were found between treatment conditions, as reflected by the squared differences of the various paris of rankings, the differences between seminar groups within treatments on all three ANOVA analyses were significant at the .01 level (Tables 4, 5, and 6). This was an unexpected finding because each seminar group was randomly assigned to one of the three treatment conditions. The results demonstrated that no one type of content rating statement was superior to any other in determining rank-order differences.

The data analyses showed that the statements investigated here yielded no significant advantages for one set of statements over another. It makes no difference whether the rating statements are task-oriented, worker-oriented, or trait-oriented. This study provides additional evidence that the doubts of Bell, Hoff, and Hoyt (1963), Borman and Dunnette (1975), and Kavanagh, MacKinney, and Wollins (1971) about the superiority of job-oriented dimensions over trait-oriented dimensions were well founded. As Kavanagh (1971) concluded from his comprehensive literature review of performance appraisal studies, there is no reason to assume the superiority of job-oriented statements over trait-oriented statements. The selection of rating statements for inclusion in performance appraisal devices should primarily be determined by cost considerations. Cost considerations tend to favor trait-oriented statements in most situations, since the job analysis required to obtain task-oriented and worker-oriented statements is costly and time consuming. Moreover, trait-oriented statements are also more generalizable across different occupations than is either task-oriented or worker-oriented statements.

Unlike many prior studies, this study does not conclude with a condemnation of judgmental rating statements. This study suggests that peer group person-oriented statements are as effective as job descriptive statements when the standard is an external criterion, such as the ability to recognize peers from unidentified profiles or the ability to predict their official class rank.

An unexpected finding was the significant effect associated with seminar groups on all performed ANOVAs, particularly since all seminar groups were randomly assigned to each treatment condition. The importance of recognizing and controlling for group effects in such performance evaluation studies is evident. Investigated treatment variables might easily become contaminated by group effects leading to inaccurate results and conclusions. The reasons for these significant group effects are unknown, although such intra-group variables as morale, leadership, and attitude are possible causal influences.

It may be that performance appraisal research emphasis has not been placed on the most important variables. Perhaps there are environmental influences that affect performance ratings more than variables attributable to the appraisal device. Perhaps such issues as content, format, scale, etc., are relatively unimportant as compared to these other variables. A need also exists to broaden the research focus in performance appraisal studies focusing on criteria independent and external to the performance appraisal device.

IV. SUMMARY AND CONCLUSIONS

Three different kinds of rating stimulus statements, differing along a dimension of trait-oriented to task-oriented descriptions, were compared in a context which permitted the comparisons to be made in terms of criteria external to the ratings. No evidence of superiority was found for any of the three sets, although many significant correlations with various external criteria were obtained in all three experimental conditions.

Significant differences were also found among the three rating sub-groups comprising each of the three treatment groups although these rating sub-groups were assigned randomly to the three treatment groups. The importance of controlling for group effects in peer group studies was noted.

REFERENCES

Arvey, R.D., & Hoyle, J.C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. *Journal of Applied Psychology*, 1974, **59**, 61–68.

Barrett, R.S. *Performance rating.* Chicago: Science Research Associates, 1966.

Bell, F.O., Hoff, A.L., & Hoyt, K.B. A comparison of three approaches to criterion measurement. *Journal of Applied Psychology*, 1963, **47**, 416–418.

Borman, W.C., & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, 1975, **60**, 561–565.

Borman, W.C., & Vallon, W.R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology*, 1974, **59**, 197–206.

Burnaska, R.F., & Hollmann, T.D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology*, 1974, **59**, 307–312.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

Campbell, J.P., Dunnette, M.D., Arvey, R.D., & Hellervik, L.V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, **57**, 15–22.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Curton, E.D., Ratliff, F.R., & Mullins, C.J. Content analysis of rating criteria. *Proceedings of Symposium on Criterion Development for Job Performance Evaluation*, 23–24 June 1977.

Dickinson, T.L., & Tice, T.E. A multitrait-multimethod analysis of scales developed by retranslation. *Organizational Behavior and Human Performance*, 1973, **9**, 421–438.

Dunnette, M.D. *Personnel selection and placement.* Belmont, CA: Wadsworth, 1966.

Ferguson, G.A. *Statistical analysis in psychology and education.* New York: McGraw-Hill, 1966.

Hays, W.L. *Statistics for psychologists.* New York: Holt, Rinehart, and Winston, 1963.

Kavanagh, M.J. The content issue in performance appraisal: A review. *Personnel Psychology*, 1971, **24**, 653–668.

Kavanagh, M.J., MacKinney, A.C., & Wollins, L. Issues in managerial performance: Multitrait-Multimethod analysis of ratings. *Psychological Bulletin*, 1971, **75**, 34–49.

Kelley, E.L., & Fiske, D.W. *The prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press, 1951.

Landy, F.J., Farr, J.L., Saal, F.E., & Freytag, W.R. Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, 1976, **61**, 750–758.

McCormick, E.J. Application of job analysis to indirect validity. *Personnel Psychology*, 1959, **12**, 402–413.

Ronan, W.E., & Prien, E.P. *Perspectives on the measurement of human performance.* New York: Appleton Century Crofts, 1971.

Schmidt, F.L., & Kaplan, L.B. Composite versus multiple criteria: A review and a resolution of the controversy. *Personnel Psychology*, 1971, **24**, 419–484.

Smith, P.C., & Kendall, L.M. Retranslation of expectations: Approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

Timm, N.H., & Carlson, J.E. *Analysis of variance through full rank models.* Multivariate Behavioral Research Monograph No. 75-1. Published by the Society of Multivariate Experimental Psychology, 1975.

Winer, B.J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

Zedeck, S., & Baker, H.T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 1972, **7**, 457–466.

Zedeck, S., & Blood, M.R. *Foundations of behavioral science research in organizations.* Monterey, CA: Brooks/Cole, 1974.

APPENDIX A: RATING DIMENSIONS

WORKER-ORIENTED RATING DIMENSIONS

	Below Average	Average	Above Average	Well Above Average	Out-Average
1. Military appearance.....	(A)	(B)	(C)	(D)	(E)
2. Participates in class activities.....	(A)	(B)	(C)	(D)	(E)
3. Communicates clearly by oral and written methods...	(A)	(B)	(C)	(D)	(E)
4. Amount of assistance to peers in work assignments..	(A)	(B)	(C)	(D)	(E)
5. Completes work in a timely manner.....	(A)	(B)	(C)	(D)	(E)
6. Follows provided instructions.....	(A)	(B)	(C)	(D)	(E)
7. Takes accurate notes.....	(A)	(B)	(C)	(D)	(E)
8. Competence in analyzing work assignments.....	(A)	(B)	(C)	(D)	(E)
9. Awareness of safety precautions.....	(A)	(B)	(C)	(D)	(E)
10. Studies well on his own...	(A)	(B)	(C)	(D)	(E)

TASK-ORIENTED RATING DIMENSIONS

	Below Average Effectiveness	Average Effectiveness	Above Average Effectiveness	Well Above Average Effectiveness	Outstanding Effectiveness
1. Knows UCMJ programmed text.....	(A)	(B)	(C)	(D)	(E)
2. Contributes examples in seminar on Discipline and Unity of Command.....	(A)	(B)	(C)	(D)	(E)
3. Promotes and organizes Community Project.....	(A)	(B)	(C)	(D)	(E)
4. Analyzes courts-martial case study.	(A)	(B)	(C)	(D)	(E)
5. Participates in Foreign Policy role playing.....	(A)	(B)	(C)	(D)	(E)
6. Understands reasons for nonalignment of uncommitted nations.	(A)	(B)	(C)	(D)	(E)
7. Knows history of AF uniform.....	(A)	(B)	(C)	(D)	(E)
8. Applies the six-step approach to problem solving.....	(A)	(B)	(C)	(D)	(E)
9. Knows how to plan a conference.....	(A)	(B)	(C)	(D)	(E)
10. Researches topic for Persuasive Speech..	(A)	(B)	(C)	(D)	(E)

TRAIT-ORIENTED RATING DIMENSIONS

	Below Average	Average	Above Average	Well Above Average	Out-standing
1. Honesty - straightforward and truthful in dealing with others.....	(A)	(B)	(C)	(D)	(E)
2. Ambition - works hard, accepts challenges.....	(A)	(B)	(C)	(D)	(E)
3. Dependability - does assigned tasks con-scientiously without close supervision.....	(A)	(B)	(C)	(D)	(E)
4. Punctuality - prompt in keeping engagements...	(A)	(B)	(C)	(D)	(E)
5. Quality of work - per-forms work accurately and effectively.....	(A)	(B)	(C)	(D)	(E)
6. Quantity of work - produces a large amount of work that meets requirement standards....	(A)	(B)	(C)	(D)	(E)
7. Initiative - originates and achieves goals on his own.....	(A)	(B)	(C)	(D)	(E)
8. Adaptability - changes attitude and behavior to meet the demands of the situation.....	(A)	(B)	(C)	(D)	(E)
9. Originality - creative, thinks of new solutions to old problems.....	(A)	(B)	(C)	(D)	(E)
10. Agreeableness - gets along well with fellow workers, well liked.....	(A)	(B)	(C)	(D)	(E)

☆U.S. GOVERNMENT PRINTING OFFICE:1979- 671-056/109